D156–D159 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh015

The KNOTTIN website and database: a new information system dedicated to the knottin scaffold

Jean-Christophe Gelly, Jérôme Gracy, Quentin Kaas[1], Dung Le-Nguyen[2], Annie Heitz and Laurent Chiche*

Centre de Biochimie Structurale, UMR 5048 CNRS INSERM Université Montpellier I, Faculté de Pharmacie, 15 avenue Charles Flahault, F-34093 Montpellier, France, [1]Laboratoire d'ImmunoGénétique Moleculaire, Université Montpellier II, UPR CNRS 1142 IGH, 141 rue de la Cardonille, F-34396 Montpellier, France and [2]U376 INSERM, Bâtiment INSERM, CHU Arnaud de Villeneuve, 371 rue du doyen Gaston Giraud, F-34295 Montpellier, Cedex 5, France

Received July 17, 2003; Revised August 19, 2003; Accepted September 3, 2003

ABSTRACT

The KNOTTIN website and database organize information about knottins or inhibitor cystine knots, small disulfide-rich proteins with a knotted topology. Thanks to their small size and high stability, knottins provide appealing scaffolds for protein engineering and drug design. Static pages present the main historical and recent results about knottin discoveries, sequences, structures, folding, functions, applications and bibliography. Database searches provide dynamically generated tabular reports or sequence alignments for knottin three-dimensional structures or sequences. BLAST/HMM searches are also available. A simple nomenclature, based on loop lengths between cysteines, is proposed and is complemented by a uniform numbering scheme. This standardization is applied to all knottin structures in the database, facilitating comparisons. Renumbered and structurally fitted knottin PDB files are available for download. The standardized numbering is used for automatic drawing of two-dimensional Colliers de Perles. The KNOTTIN website and database are available at http://knottin.cbs.cnrs.fr and http://knottin.com.

SMALL DISULFIDE-RICH PROTEINS WITH A KNOTTED ARRANGEMENT

The elucidation, in 1982, of the X-ray structure of PCI, a carboxypeptidase inhibitor from potato, revealed for the first time a 'knotted' topology in which one disulfide bridge was shown to penetrate a macrocycle formed by two other disulfides and the interconnecting backbone segments (1). In 1989, this peculiar scaffold was shown to also appear in the squash trypsin inhibitors (2–4), and, later on, in toxins from cone snails and spiders (5,6). This structural scaffold has now been found in 12 different protein families and more than 80 experimentally determined structures. We proposed that this structural family be referred to as knottins (7), although other names were later suggested, i.e. inhibitor cystine knots (8). The specific interest in this particular scaffold has come from the observation that these proteins are very small, and thus readily accessible to chemical synthesis, yet remarkably stable thanks to the high content of disulfide bridges and the 'knotted' topology. Various uses of this scaffold have been reported in protein engineering, drug design and combinatorial approaches (9–13), and reviews have been published (14,15).

THE KNOTTIN WEBSITE

Despite the strong potential of the knottin scaffold, very little is known of the sequence-structure relationships in knottins since, besides cysteines, virtually no sequence conservation is observed between families. Moreover, these proteins lack a large hydrophobic core and standard secondary structures, and a major part of their stability comes from the disulfide links. All this renders rational design and stability predictions difficult. It is therefore of interest to gather all information on knottins in one place to assist in the better understanding of sequence–structure–function relationships.

With this in mind, we have set up a dedicated information system, the KNOTTIN website, which gathers essential data on knottin discoveries, folding, applications, functions and bibliography. This is complemented by the KNOTTIN database, a relational database that stores information on known structures and sequences. Essential data are automatically extracted from the Protein Data Bank (16) and the SwissProt databank (17). Then, the new knottin nomenclature and numbering (see below) are computed and stored in the database as well as additional geometrical data (secondary structures, hydrogen bonds, contacts, solvent accessibilities, etc.) and schematic drawings. The KNOTTIN website is freely available at http://knottin.cbs.cnrs.fr or http://knottin.com.

KNOTTIN NOMENCLATURE, UNIQUE NUMBERING AND COLLIERS DE PERLES

To facilitate analyses and comparisons, a new nomenclature and a unique numbering scheme are proposed and applied

*To whom correspondence should be addressed. Tel: +33 4 67 04 34 32; Fax: +33 4 67 52 96 23; Email: chiche@cbs.cnrs.fr

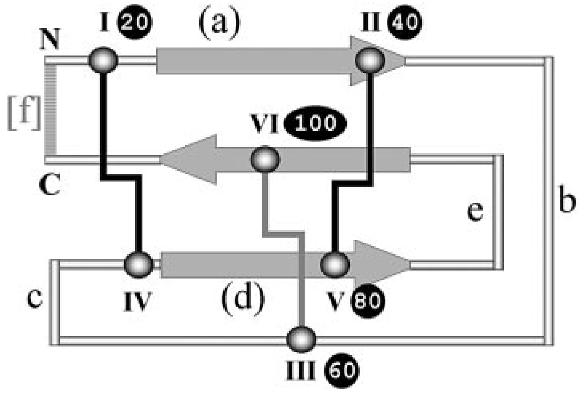

Figure 1. Schematic representation of knottins with an indication of nomenclature and unique numbering. The arrows indicate the triple-stranded β-sheet present in many knottins. The cysteines involved in the knot are labeled with roman numbers according to the sequence order. The letters a–f indicate the loop lengths between cysteines and are used to build the nomenclature. The C–N link in macrocyclic knottins is shown as a dashed line and is labeled with a gray letter. The new cysteine numbering is shown as white numbers on black circles.

throughout the structural database. The knottin scaffold is based upon the I–IV, II–V, III–VI connectivity of six cysteines to form three disulfide bridges (Fig. 1).

The proposed nomenclature indicates successively the lengths of the loops between cysteines I and II, II and III, etc., shown by a–e labels in Figure 1. The two loops involved in the disulfide macrocycle are shown in parentheses, and if necessary, numbers are separated by dots [example of nomenclature for PDB ID 2eti: '(6)5.3(1)5']. For macrocyclic knottins, in which cysteines VI and I are connected by a peptidic segment, an additional loop length is shown in brackets {example of nomenclature for macrocyclic PDB ID 1ha9: '(6)5.3(1)5[8]'}. It is worth noting that this nomenclature could easily be generalized to the growth factor cystine knots, the only other structural protein family with a disulfide bridge penetrating a disulfide macrocycle [possible nomenclature for the growth factor cystine knot PDB ID 1bet: '42(9)11.27(1)']. Note that the positions of the parentheses would simply distinguish between knottins and growth factor cystine knots. Growth factor cystine knots are not currently included in the KNOTTIN database. Additionally, a uniform numbering system has been set up for all knottins, whatever their function or origin. This greatly facilitates sequence and structure comparisons between structurally similar but sequentially divergent knottins. Such a unique numbering has already proved extremely useful for immunoglobulins and T cell receptors (18). The knottin unique numbering is based on (i) the observed loop lengths in known knottins and the need for future insertions, (ii) the position of cysteine IV which varies between families, and (iii) the wish for a simple, easy to remember numbering. According to these criteria, knottins are renumbered as follows: cysteine I → 20, cysteine II → 40, cysteine III → 60, cysteine V → 80 and cysteine VI→ 100 (Fig. 1). Gaps are inserted in the center of the loops. Cysteine IV is numbered 61 in most knottins (cysteines III and IV are adjacent), or 77 or 78 in plant cyclotides, carboxypeptidase A inhibitor and squash inhibitors (cysteine IV precedes cysteine V by two or three positions).

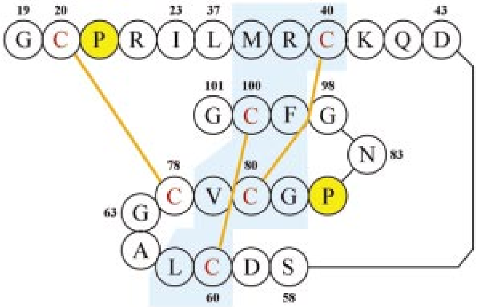

Figure 2. Two-dimensional Collier de Perles representation. Fixed residues are shown on a light blue background (residues 38–40, 60–61, 79–81 and 99–100). Their positions in the Colliers de Perles conform to the proximities in the 3D structures. As an example, cysteines 80 and 100 face each other in the β-hairpin with residues 81 and 99 being hydrogen bonded. Other residues are inserted according to the knottin numbering. Cysteines and disulfide bridges of the knot are shown as red letters and orange thick lines, respectively. Prolines are shown on a yellow background.

Thanks to the knottin unique numbering, standardized Collier de Perles representations (19) could be drawn automatically (Fig. 2). The core of the Colliers de Perles is based on the elementary Cystine Stabilized Beta-sheet (CSB) structural motif, which is the most structurally conserved part between knottins and which might correspond to a, now lost, ancestral two-disulfide scaffold (20,21).

DATABASE IMPLEMENTATION AND ACCESS

To facilitate building and updates, automatic procedures to manage the database have been written in PERL. Nevertheless, human expertise probably remains an essential component of a very specific database such as this one. The main flow chart and content are outlined in Figure 3.

Automatic detection of knottins rests mainly on (i) the I–IV, II–V, III–VI cysteine connectivity and the associated unique numbering, and (ii) the observation that the knottin scaffold is based on the structural CSB elementary motif (20). The latter is checked via structural superimposition, and the corresponding root mean square deviation (RMSD), of residues 40, 60–61, 79–81 and 99–100 onto a reference structure, i.e. the squash trypsin inhibitor CPTI–II in complex with bovin trypsin (PDB ID: 2btcI, resolution: 1.5 Å). Structures are either selected (RMSD ≤ 1.0 Å), rejected (RMSD ≥ 3.0 Å), or marked for manual inspection (1.0 Å < RMSD ≤ 3.0 Å). This procedure to automatically recognize knottin structures actually retrieved all previously known knottins. It also permitted the discovery of a new knottin member among the scorpion toxin family (chlorotoxin, PDB ID 1chl). Although scorpion toxins are based on the same elementary CSB motif as knottins (20), the disulfide connectivity is different and the third disulfide bridge between the N- and C-termini does not form a knot. However, in chlorotoxin, a fourth disulfide bridge is present with a connectivity corresponding to the I–IV bridge in other knottins. Homologs of known knottins were then searched in the SwissProt/TrEMBL database using BLAST (22) and HMMER (23) programs with low cut-offs followed by manual elimination of irrelevant hits. Cross-links between PDB IDs and SwissProt IDs were manually checked and extended when possible. Data are stored in several tables in a

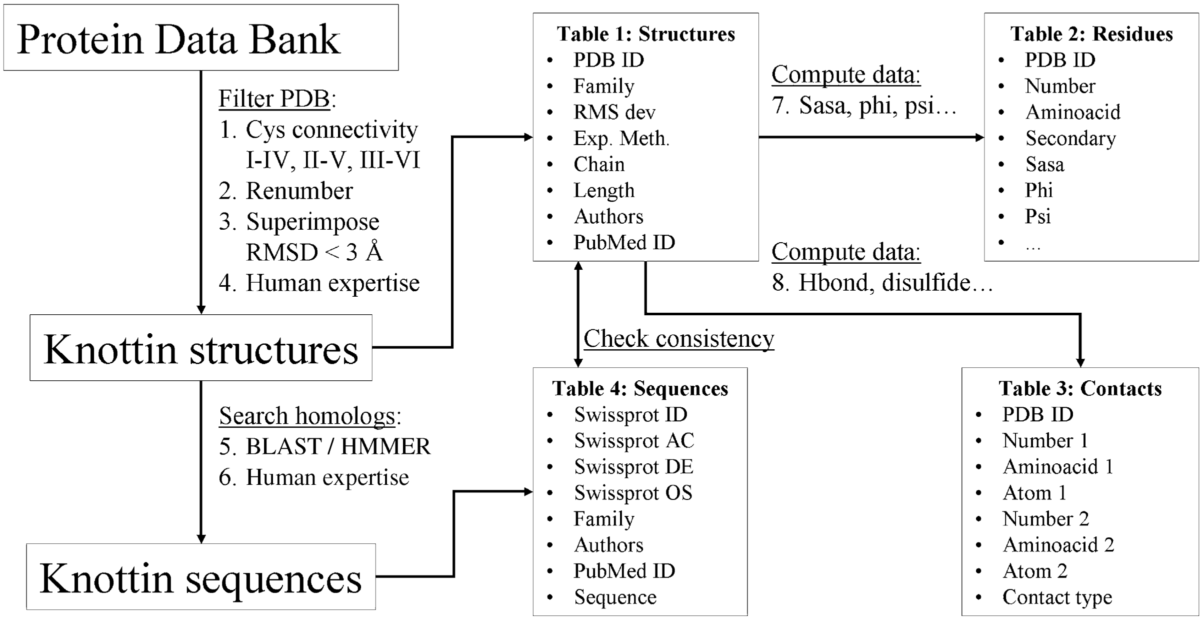

Figure 3. Main flow chart and content of the KNOTTIN database.

MySQL relational database management system. The current version of the KNOTTIN database contains 85 3D structures and 385 sequences.

Database searches can be performed through PHP or PERL scripts. Currently, users can carry out the following functions.

(i) Search sequences and/or structures for family, function, source, nomenclature, and display tabular reports or sequence alignments. The nomenclature and the RMSD from the reference structure (PDB ID: 2btc, chain I) are displayed, as well as links to SwissProt, PDBsum, MMDB and PubMed. Images of the two-dimensional Colliers de Perles are shown for each knottin.

(ii) Download PDB files renumbered according to the knottin numbering scheme, or PDB files renumbered and fitted to the structural reference.

(iii) BLAST/HMMER a sequence against the knottin database.

(iv) Search knottin structures for particular sequence or geometrical pattern (Segment search).

(v) Renumber, superimpose, establish the nomenclature and display two-dimensional representations for user-uploaded knottin structures.

Suggestions or additional data should be directed to L. Chiche at chiche@cbs.cnrs.fr, and this article should be cited when using the KNOTTIN website or database in research projects.

FUTURE DEVELOPMENTS

We plan to rapidly extend the system along several directions: (i) enrich the static pages with additional information and make the bibliography searchable through the MySQL database, (ii) add new search and display types since several data stored in the database are not currently used, i.e. hydrogen bonds and contacts, (iii) build accurate homology models for knottin sequences that lack experimental 3D structure.

ACKNOWLEDGEMENTS

We thank Jean-Luc Pons for help in web developments and Marie-Paule Lefranc, for authorizing the use of the expression ‘Collier de Perles’, which originally referred to standardized 2D representations in IMGT, the international ImMunoGenetics information system (http://imgt.cines.fr). This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Institut National de la Santé et de la Recherche Médicale (INSERM) and the Génopole Montpellier Languedoc-Roussillon.

REFERENCES

1. Rees,D.C. and Lipscomb,W.N. (1982) Refined crystal structure of the potato inhibitor complex of carboxypeptidase A at 2.5 Å resolution. *J. Mol. Biol.*, **160**, 475–498.
2. Bode,W., Greyling,H.J., Huber,R., Otlewski,J. and Wilusz,T. (1989) The refined 2.0 Å X-ray crystal structure of the complex formed between bovine β-trypsin and CMTI-I, a trypsin inhibitor from squash seeds (*Cucurbita maxima*). Topological similarity of the squash seed inhibitors with the carboxypeptidase A inhibitor from potatoes. *FEBS Lett.*, **242**, 285–292.
3. Chiche,L., Gaboriaud,C., Heitz,A., Mornon,J.P., Castro,B. and Kollman,P.A. (1989) Use of restrained molecular dynamics in water to determine three-dimensional protein structure: prediction of the three-dimensional structure of *Ecballium elaterium* trypsin inhibitor II. *Proteins*, **6**, 405–417.
4. Heitz,A., Chiche,L., Le-Nguyen,D. and Castro,B. (1989) 1H 2D NMR and distance geometry study of the folding of *Ecballium elaterium* trypsin inhibitor, a member of the squash inhibitors family. *Biochemistry*, **28**, 2392–2398.
5. Davis,J.H., Bradley,E.K., Miljanich,G.P., Nadasdi,L., Ramachandran,J. and Basus,V.J. (1993) Solution structure of ω-conotoxin GVIA using 2-D NMR spectroscopy and relaxation matrix analysis. *Biochemistry*, **32**, 7396–7405.

6. Yu,H., Rosen,M.K., Saccomano,N.A., Phillips,D., Volkmann,R.A. and Schreiber,S.L. (1993) Sequential assignment and structure determination of spider toxin ω-Aga-IVB. *Biochemistry*, **32**, 13123–13129.
7. Le-Nguyen,D., Heitz,A., Chiche,L., Castro,B., Boigegrain,R.A., Favel,A. and Coletti-Previero,M.A. (1990) Molecular recognition between serine proteases and new bioactive microproteins with a knotted structure. *Biochimie*, **72**, 431–435.
8. Pallaghy,P.K., Nielsen,K.J., Craik,D.J. and Norton,R.S. (1994) A common structural motif incorporating a cystine knot and a triple-stranded β-sheet in toxic and inhibitory polypeptides. *Protein Sci.*, **3**, 1833–1839.
9. Hilpert,K., Schneider-Mergener,J. and Ay,J. (2002) Crystallization and preliminary X-ray analysis of the complex of porcine pancreatic elastase and a hybrid squash inhibitor. *Acta Crystallogr. D*, **58**, 672–674.
10. Baggio,R., Burgstaller,P., Hale,S.P., Putney,A.R., Lane,M., Lipovsek,D., Wright,M.C., Roberts,R.W., Liu,R., Szostak,J.W. *et al.* (2002) Identification of epitope-like consensus motifs using mRNA display. *J. Mol. Recognit.*, **15**, 126–134.
11. Heitz,A., Le-Nguyen,D., Dumas,C. and Chiche,L. (2000) Engineering potential inhibitors of the interaction between the HIV-1 NEF protein and kinase SH3 domains. In Martinez,J. and Fehrentz,J.A. (eds), *Peptides 2000*. Editions EDK, Paris, France, pp. 415–416.
12. Craik,D., Daly,N.L. and Nielsen,K.J. (2000) Preparation of cyclized conotoxin peptides. PTC International Patent Application WO 0015654.
13. Smith,G.P., Patel,S.U., Windass,J.D., Thornton,J.M., Winter,G. and Griffiths,A.D. (1998) Small binding proteins selected from a combinatorial repertoire of knottins displayed on phage. *J. Mol. Biol.*, **277**, 317–332.
14. Norton,R.S. and Pallaghy,P.K. (1998) The cystine knot structure of ion channel toxins and related polypeptides. *Toxicon*, **36**, 1573–1583.
15. Craik,D.J., Daly,N.L. and Waine,C. (2001) The cystine knot motif in toxins and implications for drug design. *Toxicon*, **39**, 43–60.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
18. Lefranc,M.P., Pommie,C., Ruiz,M., Giudicelli,V., Foulquier,E., Truong,L., Thouvenin-Contet,V. and Lefranc,G. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol.*, **27**, 55–77.
19. Lefranc,M.P., Giudicelli,V., Ginestoux,C., Bodmer,J., Muller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbie,V. and Chaume,D. (1999) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **27**, 209–212.
20. Heitz,A., Le-Nguyen,D. and Chiche,L. (1999) Min-21 and min-23, the smallest peptides that fold like a cystine-stabilized β-sheet motif: design, solution structure, and thermal stability. *Biochemistry*, **38**, 10615–10625.
21. Wang,X., Connor,M., Smith,R., Maciejewski,M.W., Howden,M.E., Nicholson,G.M., Christie,M.J. and King,G.F. (2000) Discovery and characterization of a family of insecticidal neurotoxins with a rare vicinal disulfide bridge. *Nature Struct. Biol.*, **7**, 505–513.
22. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
23. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.